AI – Service Description

VERSION JUNE 2026

1NCE
Cologne – Miami – São Paulo – Singapore – Tokyo

1 Introduction

1NCE AI provides business customers with unified access to third-party generative AI models ("AI Models") through a single application programming interface (API). 1NCE AI operates as a managed inference gateway: it receives a request ("Input") from the Customer, routes it to the AI Model selected by the Customer, and returns the model's response ("Output").

1NCE AI is a pass-through service. 1NCE does not modify the behaviour of the AI Models and does not add its own AI decision logic. The AI Models are operated by third-party providers and made available through Amazon Web Services (AWS) Bedrock.

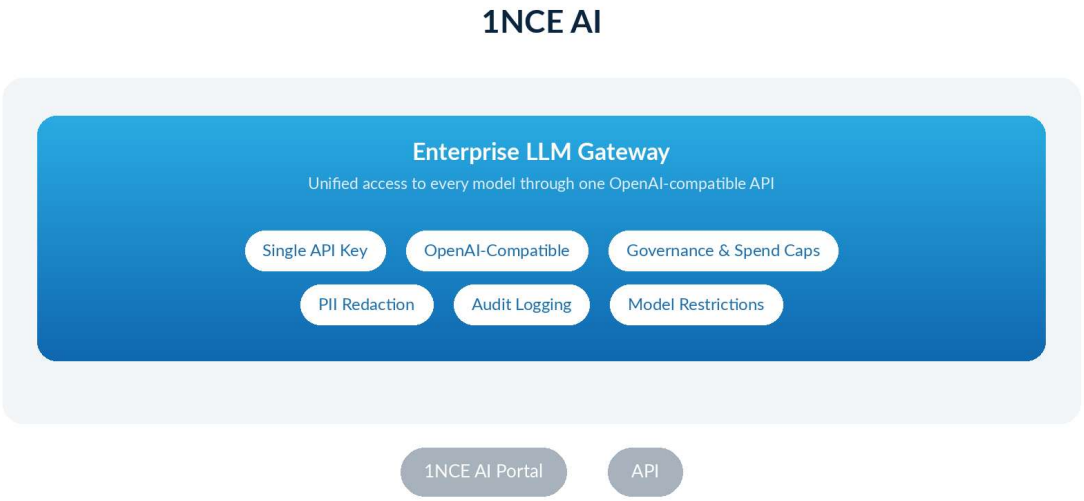
1NCE AI is provided to business customers only. Customers access 1NCE AI through the API (see Section 2.3) and manage their account through the 1NCE AI Portal (see Section 3).

This Service Description is applicable only if the Customer has an active 1NCE AI account (see Terms & Conditions).

The definitions set out in the Terms & Conditions also apply to this document.

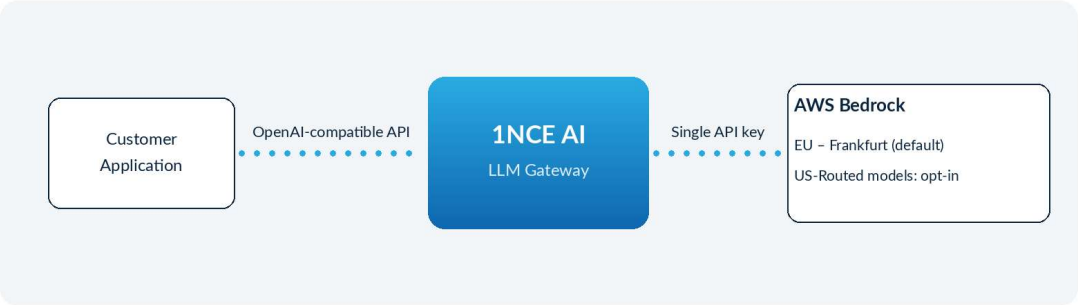
2 1NCE AI

1NCE AI consists of: an inference API that routes Inputs to AI Models and returns Outputs; a catalogue of selectable AI Models (see Section 2.2); access- and key-management functions; governance controls for usage and spend (see Section 2.4); and usage monitoring and audit logging (see Section 2.5). All functions are accessible through the API (see Section 4) and, where applicable, the 1NCE AI Portal (see Section 3).



2.1 Service Overview

When the Customer submits an Input to the API, 1NCE AI routes it to the AI Model identified in the request, receives the resulting Output from the AI Model, and returns the Output to the Customer. Inputs and Outputs are processed to fulfil the request; 1NCE AI does not use Inputs or Outputs to train, fine-tune or evaluate any AI Model.



2.2 Model Catalogue

2.2.1 Available Models

1NCE AI provides access to a range of AI Models from multiple providers through a single API key. The full, current Model Catalogue is available in the 1NCE AI Portal, after sign-in; current per-Model rates are set out on the Pricing Page at https://portal.1nce.ai/dashboard/pricing.

The Model Catalogue is maintained by 1NCE and is updated when AI Models are added, changed or withdrawn (for example, as models become available on, or are withdrawn from, AWS Bedrock). Most AI Models are processed within the European Union (AWS eu-central-1, Frankfurt). The Models identified in the Model Catalogue as "US-Routed Models" are processed in the United States and must be explicitly enabled by the Customer before use (opt-in). The data-protection treatment of such processing is addressed in the Data Processing Addendum.

2.2.2 Model Selection and Routing

The Customer selects the AI Model for each request by specifying the model identifier in the API call.

2.2.3 API Access

2.2.4 Endpoint and Compatibility

1NCE AI exposes a single inference API endpoint at https://llm.1nce.ai/v1. The API is compatible with the OpenAI API request and response format. Existing OpenAI-compatible SDKs and tools (for example LangChain, LlamaIndex and LiteLLM) can be used by configuring the base URL and API key. Streaming responses are supported in the OpenAI-compatible format.

2.2.5 Authentication and API Keys

Requests are authenticated using an API key. The Customer generates scoped API keys through the 1NCE AI Portal (see Section 3). API keys can be created, rotated and revoked by the Customer at any time. Each API key can be assigned its own spend cap, model restrictions and per-request token limit (see Section 2.4).

2.2.6 Governance and Controls

2.2.7 Customer-Configured Spend Caps and Model Restrictions

The Customer can configure, per API key, a monthly spend cap, an allowed-models list and a maximum input-token size. Platform rate limits (see Section 2.4.2) are fixed and not customer-configurable. Low-balance warnings are provided when the prepaid account balance runs low.

2.2.8 Platform Rate Limits

To maintain the stability, security and availability of the Service for all customers, 1NCE applies platform-level rate limits to API requests, independently of, and in addition to, any limits the Customer configures. These limits are applied per API key and per customer account and may include limits on the number of requests per minute and the number of tokens per minute. The applicable limits are set out in the Developer Hub (see Section 5) and may be adjusted by 1NCE where reasonably necessary to protect the Service. By default, 1NCE applies the following platform limits: 60 requests per minute and 100,000 tokens per minute per API key, with a maximum of 10 concurrent requests per key, and 200 requests per minute per customer across all API keys. Higher limits may be available on request.

2.2.9 Spend Caps and Budgets

The Customer can configure a monthly spending cap per API key. A per-key spend cap is a best-effort guardrail; the prepaid account balance is the hard ceiling. When the balance is exhausted, further requests are blocked (HTTP 402). Consumption is tracked in the 1NCE AI Portal (see Section 3).

2.2.10 PII Redaction

1NCE AI applies a server-side PII redaction step within the API request path; it is not an interactive or in-chat feature. Before an Input is transmitted to the AI Model, a detection process identifies personal-data entities (such as names, email addresses, telephone numbers, IBANs, payment-card numbers, IP addresses, and national identifiers) and replaces them with typed placeholders (for example <PERSON> or <EMAIL_ADDRESS>). The step applies to all API requests, including programmatic API-key calls, and is enabled by default; it operates on Input content before it reaches the AI Model. Redaction applies to Input only; Output is not filtered. The redaction step is fail-closed: if the redaction service is unavailable, the request is rejected rather than sent unredacted. System prompts are not redacted by default.

2.3 Usage and Audit Logging

1NCE AI records per-request metadata, including the API-key identifier, AI Model identifier, token counts, latency and response codes. Audit logs do not contain the content of Inputs or Outputs. The metadata is available to the Customer through the 1NCE AI Portal (see Section 3) and the API (see Section 4) for monitoring and reconciliation.

3 1NCE AI Portal

1NCE AI is managed through the 1NCE AI Portal at portal.1nce.ai. The 1NCE AI Portal is a dedicated, separately branded portal for the Service and is distinct from the 1NCE IoT connectivity Customer Portal. Customers register as new users; the 1NCE AI Portal provides:

Account and organisation management;

API-key generation and management (see Section 2.3.2);

Model selection and configuration (see Section 2.2);

Spend-cap and model-restriction configuration (see Section 2.4);

Usage and cost dashboards (see Section 2.5).

1NCE AI provides three roles: Admin (full access), Developer (API keys and usage) and Viewer (read-only), enforced server-side.

4 API

1NCE AI does not currently expose a separate programmatic management API. The API surface is the OpenAI-compatible inference API described in Section 2.3 (chat completions and model listing). Communication with the API is via HTTPS using JSON content. API keys, budgets and usage are managed by the Customer through the 1NCE AI Portal (see Section 3). The API documentation is available in the Developer Hub (see Section 5).

5 Developer Hub

1NCE provides developer documentation, quickstart guides and usage examples for 1NCE AI in the Developer Hub at https://start.1nce.ai/docs.

6 Support and Service Levels

Support for 1NCE AI is provided under the Standard service tier described in the 1NCE Global Service Description. No Premium support tier and no 1NCE AI-specific support model apply.

The Service is provided on a best-efforts basis. No binding service level (including availability or uptime) applies to the Service.